9

SCALING OF MOS TECHNOLOGY TO SUBMICROMETER FEATURE SIZES

Carver A. Mead *

Abstract

Industries based on MOS technology now play a prominent role in the developed and the developing world. More importantly, MOS technology drives a large proportion of innovation in many technologies. It is likely that the course of technological development depends more on the capability of MOS technology than on any other technical factor. Therefore, it is worthwhile investigating the nature and limits of future improvements to MOS fabrication. The key to improved MOS technology is reduction in feature size. Reduction in feature size, and the attendant changes in device behaviour, will shape the nature of effective uses of the technology at the system level. This paper reviews recent, and historical, data on feature scaling and device behavior, and attempts to predict the limits to this scaling. We conclude with some remarks on the system-level implications of feature size as the minimum size approaches physical limits.

9.1 Introduction

It is always difficult to predict the future; few attempts to do so have met with resounding success. One remarkable example of successful prediction is the exponential increase in complexity of integrated circuits, first noted by Gordon E. Moore. As we contemplate the ongoing evolution of this great technology, many questions arise: Can the trend continue? Will single-chip systems attain levels of complexity that render present system architectures unworkable [1]? Will digital techniques completely replace analog methods [2]? The answers to these questions depend critically on the properties of the individual transistors that provide the essential active functions, without which no interesting system behavior is possible. Integrated-circuit density is increased by a reduction in the size of elementary features of the underlying structures; therefore, any discussion of the capabilities of future technologies must rely on an understanding of how the properties of transistors evolve as the transistors' dimensions are made smaller.

Elsewhere [3], we described the factors that limit how small an MOS transistor

*Reproduced from Journal of VLSI Signal Processing, **8**, 9-25 (1994) Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

can be and still operate properly. That discussion will not be repeated here, but I will outline the major issues:

1. For the device current to be primarily controlled by the gate, the device should not be punched through; that is, the sum of the source and drain depletion layers should be less than the geometric channel length. As a direct consequence of this requirement, the bulk doping must increase as dimensions are decreased.

2. Increasing the bulk doping has two important consequences:

   a. Junction breakdown voltage is lowered.

   b. A larger electric field is required in the gate oxide to obtain a given change in surface potential.

Because of 2a, the operating voltage must be reduced. So that sufficient electric field can be obtained with a lower operating voltage, the gate oxide must be made thinner. Thus, it is inevitable that, as the minification process is continued, both drain depletion layer and gate oxide will become thin enough that electron tunneling through them will become comparable with other device currents. In 1971, when our original study [3] was written, we described a device of 0.15 micrometer (μ) channel length, having a 50 Angstrom ($\mathring{A}$) gate oxide. Although we were confident that a device of this size could be made to work, we were not at all sure that smaller devices could be made viable.

Over the ensuing 22 years, feature sizes have evolved from 6 to 0.6 μ and the trend shows no sign of abating [4–10]. In this paper, I shall examine what we have learned from the past 22 years of technology evolution, and shall discuss to what extent these same trends may continue into the future. I shall conclude that we can safely count on at least one more order of magnitude of scaling, with a concomitant increase in both density and performance. Several of the conclusions of this study were reached independently by Hu [11].

9.2 Scaling Approach

In Figure 9.1, I have plotted the historic trend of gate-oxide thickness t_{ox} as a function of l, the minimum feature size of the process. The trend can be expressed accurately as

$$t_{ox} = 210l^{0.77}$$

where the feature size is in μ, and the gate-oxide thickness is in $\mathring{A}$. This observation suggests that it may be fruitful to express all important process parameters as

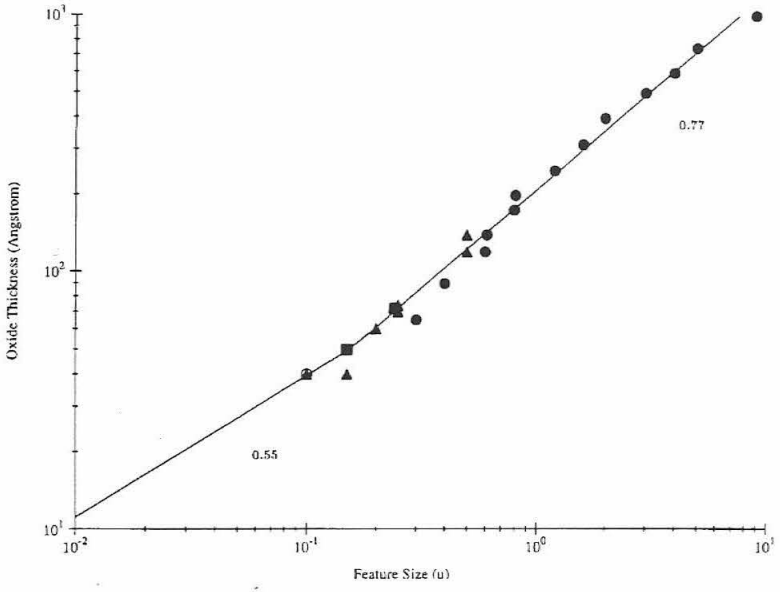

Fig. 9.1. Gate-oxide thickness as a function of feature size. The solid circles are production processes in silicon-gate technology, starting in 1970. Triangles are processes reported in the literature. Solid squares are the two most advanced devices described in our previous study [3]. The solid line is the analytic expression used in this study (Equation 9.1).

powers of the feature size, and to determine whether there is a scaling of this form that allows sensible process evolution to dimensions well below 0.1 μ. To prevent the gate oxide thickness from becoming thinner than a single atomic layer, I have chosen a scaling of the form

$$t_{ox} = \max(210l^{0.77}, 140l^{0.55}) \tag{9.1}$$

This expression is plotted as the solid line in Figure 9.1. In reviewing the historic trend, it is clear that we expressed previously [3] more concern with gate-oxide tunneling than has been justified by the experience accumulated through the intervening years. It is conceivable that I am repeating the same bit of paranoia here. In any case, if oxide thickness continues to decrease at the present rate, the resulting devices will be somewhat more capable than those I present.

The oxide thickness and feature size together determine the gate-oxide capacitance C_g of a minimum-sized device:

$$C_g = \epsilon_{ox} \frac{l^2}{t_{ox}}$$

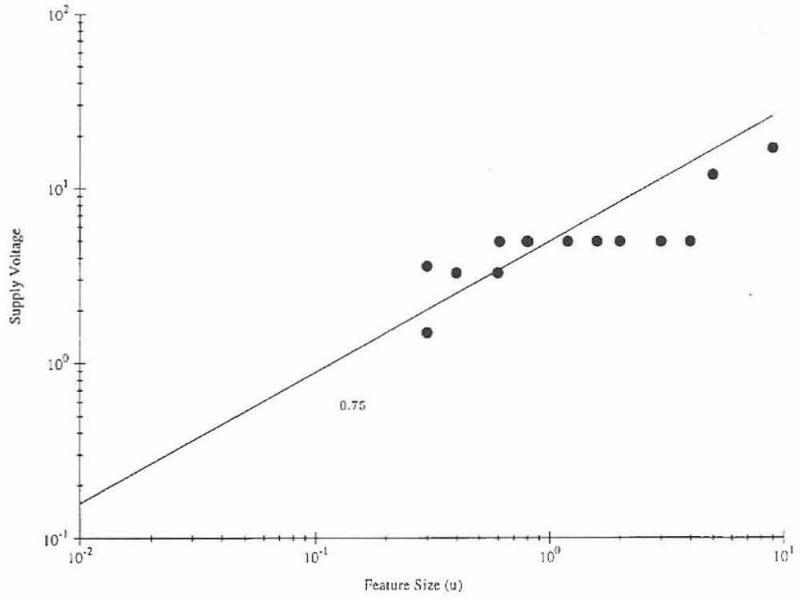

Fig. 9.2. Power-supply voltage as a function of feature size. The solid line is the analytic expression used in this study (Equation 9.2).

The historic trend in supply voltage V is shown in Figure 9.2. This trend is not as smooth as the trend in oxide thickness, due to the long period of standardization at 5 volts (V). It is clear, however, that modern submicrometer devices operate better on lower voltages [7, 12], and that this trend to lower voltages must continue. The scaling I use in this study is

$$V = 5l^{0.75} \tag{9.2}$$

This expression is plotted as the solid line in Figure 9.2.

Once we have the gate-oxide capacitance and supply voltage, we can estimate the energy W_g stored on the gate of a minimum-sized transistor at any given feature size. I have slightly overestimated the stored energy as

$$W_g = \frac{1}{2} C_g V^2 \tag{9.3}$$

For the scaling laws given here, the stored energy (in Joules) works out to be

$$W_g = 2.2 \times 10^{-14} l^{2.75} \tag{9.4}$$

This expression is plotted as the long solid line in Figure 9.3. Even with the slight "kink" introduced by Equation 9.1, this expression is a good abstraction of

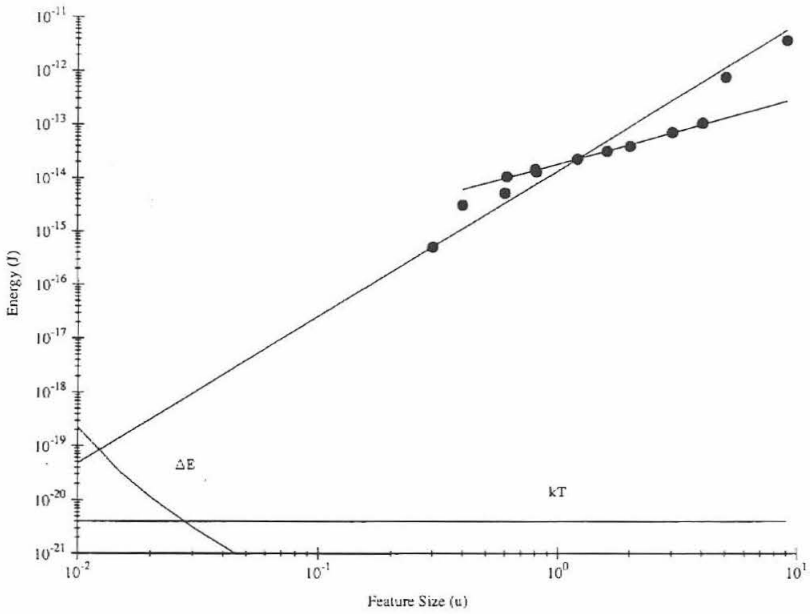

Fig. 9.3. Energy stored on the gate of a minimum-sized transistor as a function of feature size. We compute the points from Equation 9.3 using oxide-thickness values from figure 1 and the supply-voltage values from figure 2. The solid line is the analytic expression used in this study (Equation 9.4). Also shown for reference are the thermal energy kT at room temperature, and the quantum-level spacing for electrons in the channel with momenta in the direction of current flow.

the actual energy over the entire range of the plot. In the central section of historic data, however, the constant 5-V power-supply voltage has established a trend with much less dependence on feature size.

This shorter trend is well represented by the expression

$$W_5 \times 2 \times 10^{-14} l^{1.22} \tag{9.5}$$

Also shown for reference on Figure 9.3 is the thermal energy kT, and the spacing of levels in the channel with momenta in the direction of current flow. It is clear that the stored energy is more than 10 kT even at feature sizes of 0.01μ.

The minimum stored energy is an interesting quantity because it sets the scale for the switching energy dissipated in a digital system. The energy per operation of computation-intensive digital chips is compared with the minimum stored energy in Figure 9.4. The system energy per operation is four to six orders of magnitude higher than the minimum stored energy, and can be bounded by the two solid trend

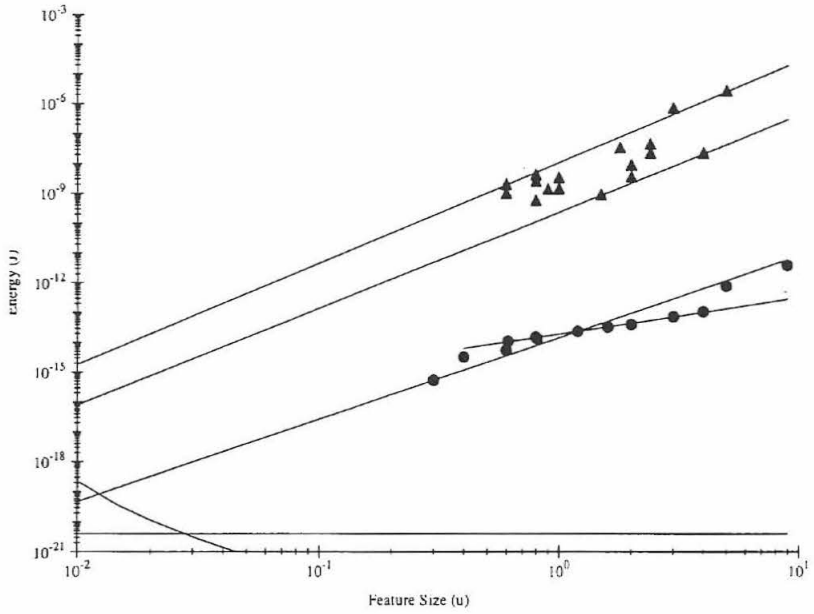

Fig. 9.4. Energy dissipated per operation at the chip level. Filled circles are data taken from the literature and from manufacturers' data sheets. Examples are all computation-intensive single chips, such as multipliers, digital signal processors, and similar devices. So that the data could be plotted on a single scale, all values were normalized to 8 x 8 multiply-add operations, assuming that the energy is proportional to the product of the word lengths of the multiplicand and multiplier. Minimum and maximum trend lines shown are Equations 9.5 and 9.6. Also shown for reference are the data of Figure 9.3.

lines:

$$W_{\max} = 1.15 \times 10^{-8} l^{3.4} \tag{9.6}$$

$$W_{\min} = 2.5 \times 10^{-10} l^{3.25} \tag{9.7}$$

The overall system trend is steeper than that for minimum stored energy, presumably because designers have become more skilled over the years, and processes have an ever increasing set of features on which designers can draw (multiple levels of metal, for example). A 5-V subtrend is clearly discernible in the system data as well.

With the information on hand, we can determine the tunneling current density J_{ox} through the gate oxide [13–15], making the worst-case assumption that the

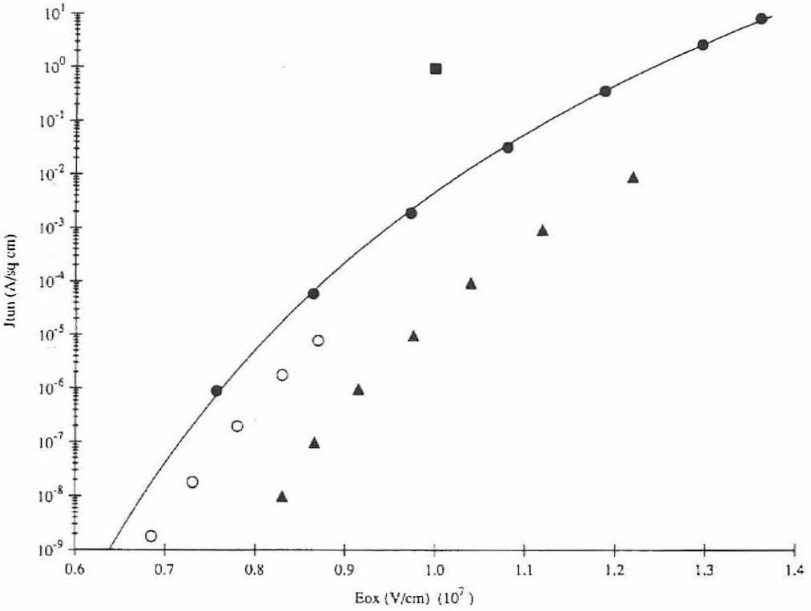

Fig. 9.5. Oxide tunneling current as a function of electric field. The open circles are from the original work of Lenzlinger and Snow [13]. Filled circles are from the recent work of Suñé et al. [15]. Filled triangles are from Hori et al. [14]. The solid line is the analytical expression used in this study (Equation 9.7). The filled square is inferred from Iwase et al. [10], but is not directly comparable with the other data because it was taken from a transistor drain characteristic, and may be corrupted with other effects such as gate-enhanced drain tunneling. The gate current was not reported separately, so this value shown represents a worst-case estimate.

entire supply voltage appears across the entire gate area:

$$J_{ox} = J_0 E_{ox}^2 e^{-kt_{ox}} \tag{9.8}$$

where $J_0 = 6.5 \times 10^{10}$ A/V/cm^2 was adjusted to match experimental data, as shown in Figure 9.5. The imaginary part of the wave vector k is given by

$$k = \frac{2k_0}{3}\frac{\phi}{V}\left[1 - \left(1 - \min\left(1, \frac{V}{\phi}\right)\right)^{3/2}\right] \tag{9.9}$$

These expressions are valid for voltages both above and below the barrier potential ϕ which was taken to be 3.2 V. The preexponential constant $k_0 = 1.2$ $\mathring{A}^{-1}$ was used. It is comforting to note that oxide tunneling data are available over the entire range of electric fields that will be encountered down to the smallest dimensions

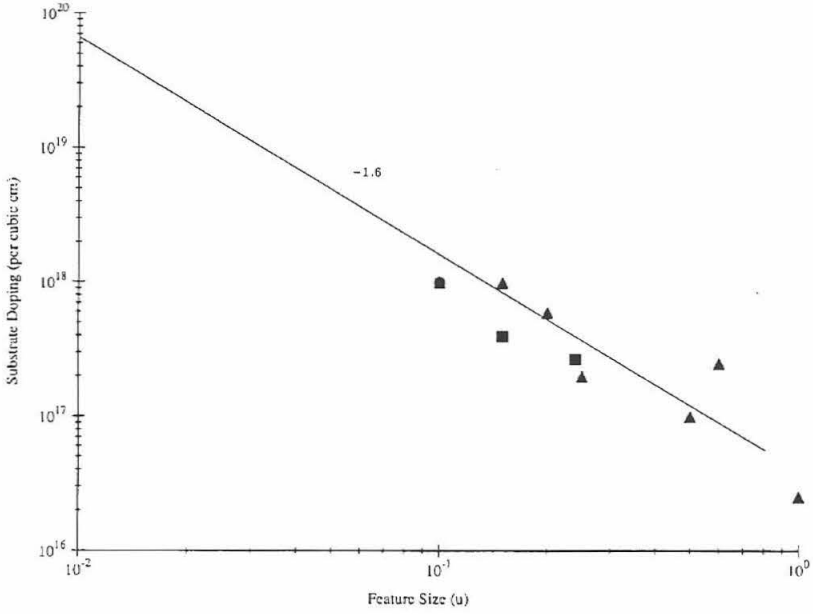

Fig. 9.6. Substrate doping as a function of feature size. The solid line is the analytical expression used in this study (Equation 9.8). Filled triangles represent processes reported in the literature. The two solid squares are the two smallest transistor designs shown in our earlier work [3].

studied here. It will be helpful, however to have actual experimental data in the 10 $\mathring{A}$ range. For these extremely thin oxides, it will be essential to take into account the quantum corrections discussed in Suñé et al. [15].

The other major source of parasitic current is tunneling through the drain junction. The junction-tunneling current density J_j is critically dependent on the substrate acceptor concentration n, which must be increased to avoid punch-through as device dimensions are decreased [16–22]. The scaling law used in this study is plotted in Figure 9.6:

$$n = 4 \times 10^{16} l^{-1.6} \tag{9.10}$$

Given the doping density n, we can compute the depletion-layer thickness x for any potential ψ relative to substrate using the usual step-junction approximation:

$$x = \sqrt{\frac{2\epsilon_{si}\psi}{qn}} \tag{9.11}$$

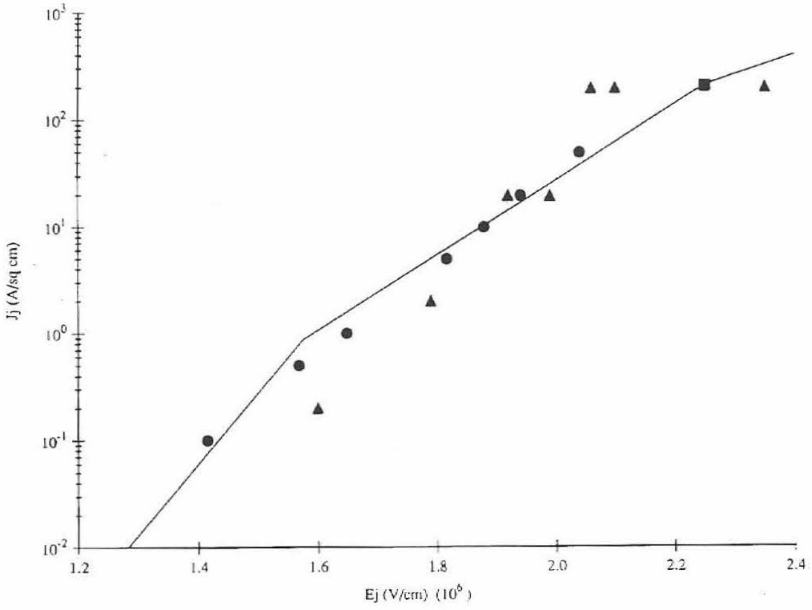

Fig. 9.7. Junction-tunneling current density as a function of peak electric field in the junction. The filled triangles are from alloy tunnel diodes, which were reported as step junctions by Chynoweth et al. [16]. The filled circles are from diffused emitter-base junctions reported as graded junctions by Fair and Wivell [19]. These were the only references that I was able to locate for electric fields in the range encountered in the finest feature sizes considered in this study. Some data are shown by Reisch [22], but not enough information is given to allow direct comparison with the other data. For reference, the solid square represents the parameters encountered in the 0.03-μ device described in this study. The solid line is the analytical expression used in this study (Equation 9.10).

The corresponding depletion-layer capacitance C is given by

$$C = \frac{\epsilon_{si}}{x}$$

We can determine the maximum electric field in the drain junction, from the junction voltage, which in the worst case will be the supply voltage plus the built-in voltage:

$$E_j = \sqrt{\frac{2qn(V + V_b)}{\epsilon_{si}}}$$

We could alternatively use a graded-junction approximation, such as that used by Fair and Wivell [19]. For our purposes, the two approaches are nearly equivalent, so

I have used the simpler step-junction expression with the junction built-in voltage V_b = 1.1 V. In either case, the tunneling current density is a function of the maximum electric field:

$$J_j = G_0 V \frac{E_j}{E_0} e^{-E_0/E_j} \tag{9.12}$$

The constant E_0 = 2.9 x 10^7 V/cm was taken from Fair and Wivell [19], and the preexponential factor G0 = 3 x 10^9 A/V cm^2 was chosen to fit the experimental data plotted in Figure 9.7. It is significant that experimental data exist that allow us to predict the tunneling currents in junctions of devices down to 0.03-μ feature sizes. Previously [3], we pointed out that the "drain corner" tunneling occurs at lower voltage than that across the junction area, a fact that has received considerable attention [23]. For the present study, I will use Equation 9.10 for area tunneling, both for simplicity and because I expect considerable cleverness on the part of process designers as this phenomenon becomes limiting. Caution, however, that corner effects may significantly increase the drain tunneling over the values shown in the following Figures.

9.3 Threshold Scaling

To determine the detailed properties of small devices, we must take into account the short-channel properties, most notable of which are carrier-velocity saturation and drain-induced barrier lowering (the precursor to punch-through). Previously [2], we developed a model that gives closed-form expressions for the current in short-channel devices, including the effects of velocity saturation. To apply the model, we need some abstraction of the vertical doping profile under the gate. The most widely used such abstraction is the threshold voltage V_t. We therefore proceed by choosing a nominal threshold voltage of the form

$$V_t = 0.55 l^{0.23} \tag{9.13}$$

The actual threshold voltage will be lower than the nominal one by the amount of drain-induced barrier lowering (DIBL) [24–27]. In this study, I use the expression given by Fjeldly and Shur [28]:

$$\text{DIBL} = V \frac{x_c}{\lambda} \frac{\sinh(x_s/\lambda}{\cosh((l - x_d)/\lambda) - \cosh(x_s/\lambda)} \tag{9.14}$$

where x_s and x_d are the classical depletion-layer thicknesses of the source and drain junctions. I have used a surface potential of 0.5 V in Equation 9.9 to compute x_c, the thickness of the depletion layer under the channel. The distance scale λ is given by

$$\lambda = x_c \left(1 + \frac{C_{ox}}{C - C}\right)^{-1/2}$$

where the depletion-layer capacitance per unit area C_c from channel to substrate is

$$C_c = \frac{\epsilon_{si}}{x_c}$$

and the oxide capacitance per unit area C_{ox} from gate to channel is

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$$

The nominal threshold voltage; the actual threshold voltage, including DIBL; and the supply voltage are plotted as a function of feature size in Figure 9.8. For the scaling parameters used in this study, DIBL does not become a serious problem until feature sizes are less than 0.03 μ.

9.4 Device Characteristics

Threshold is defined as the gate voltage at which mobile charge Q_s at the source end of the channel changes the surface potential by kT/q [2]. The channel charge at threshold is

$$Q_t = \frac{kt}{q}(C_{ox} + C_c) \tag{9.15}$$

For higher gate voltages, essentially all charge on the gate attracts equal and opposite countercharge of mobile carriers in the channel. Thus, we can form an excellent estimate of the channel charge Q_s at the source end of the channel:

$$Q_s = C_{ox}(V - V_t) \tag{9.16}$$

For gate voltages below V_t, channel current decreases exponentially with decreasing gate voltage. At zero gate voltage, the channel charge is:

$$Q_s = Q_t e^{-q\kappa V_t/kT} \tag{9.17}$$

where

$$\kappa = \frac{C_{ox}}{C_c + C_{ox}}$$

Given Q_t and Q_s, we can compute the saturated channel current for a minimum-sized transistor of any given channel length using Equation (B.28) from [2]:

$$I_{sat} = Q_s \nu_0 + Q_t \nu_0 \left(\frac{l}{l_0} + 1\right) \left(1 - \sqrt{1 + \frac{2Q_s l}{Q_t l_0}\left(\frac{l}{l_0} + 1\right)^{-2}}\right) \tag{9.18}$$

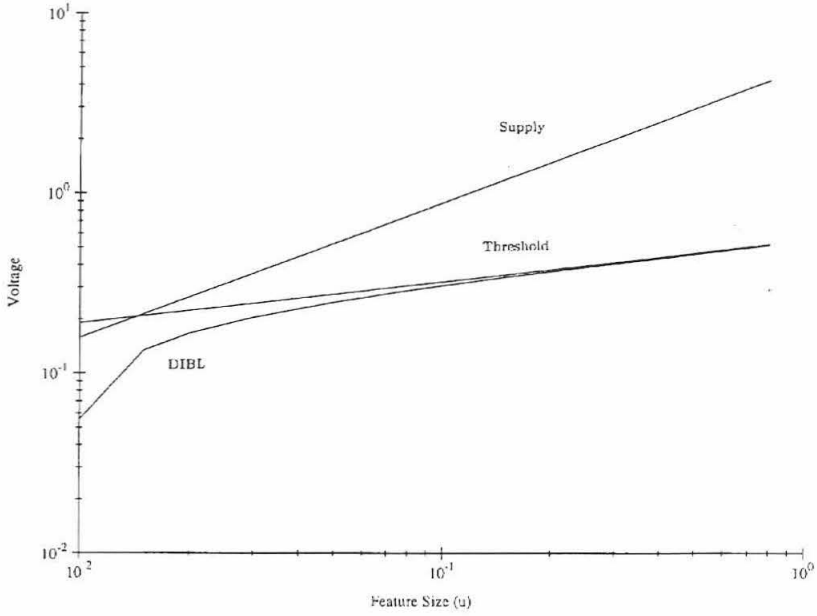

Fig. 9.8. Threshold voltage used in this study. The middle curve is the nominal threshold voltage, given by Equation 9.11. The bottom curve is the actual threshold voltage, which is lowered from the nominal value by drain-induced barrier lowering (DIBL), given by Equation 9.12. The top curve is the nominal supply voltage from Equation 9.2.

where ν_0, the saturated velocity of electrons in silicon, is taken to be 10^7 cm/s [29], and $l_0 = D/\nu_0$ can be thought of as the mean free path of the carrier, which is taken to be 0.007 μ [2].

We obtain the threshold current I_t by substituting $Q_s = Q_t$ from Equation 9.13 into Equation 9.16. We obtain the on current I_{on} by substituting Q_s from Equation 9.14 into Equation 9.16, using the threshold voltage lowered only by the built-in junction voltage, rather than by the total junction voltage. We obtain the off current (15) I_{off} by substituting Q_s from Equation 9.15 into Equation 9.16, using the threshold voltage as lowered by DIBL. These expressions thus represent a conservative characterization of the transistor performance, since the on current will be somewhat underestimated.

The several currents associated with a minimum-sized transistor are shown as a function of feature size in Figure 9.9. The trade-offs mentioned in the introduction are immediately apparent in this plot. As features become smaller, substrate doping must increase to prevent punch-through. The increase in substrate doping increases the junction electric field, thereby increasing drain-junction tunneling current into

the substrate. To limit the tunneling current to a reasonable value, we reduce the supply voltage, thereby reducing the ratio of channel on current to channel off current. The most remarkable conclusion from Figure 9.9 is that transistors of 0.03-μ channel length still function essentially as do present-day devices. With proper scaling of all parameters of the process, device miniaturization is alive and well. Many issues will arise in the development of ever-finer-scale fabrication, but, in the end, the endeavor will prevail.

Given that devices at least one order of magnitude smaller than today's are feasible, we may enquire what their characteristics may be. Figure 9.10 shows several quantities of interest. It is clear that discreteness of all quantities will become increasingly important at smaller feature sizes — particularly that of doping ions in the substrate. We have given elsewhere a simple discussion of the effects of discrete substrate charge [3]; a recent analysis is presented by Nishinohara et al. [30].

Perhaps the single most important aspect of device performance is the speed of logic fabricated from any particular technology. We can estimate the time τ required for an elementary logic element to drive another like it:

$$\tau = \frac{VC_{\text{tot}}}{I_{\text{on}}} \tag{9.19}$$

where the total capacitance C_{tot} is taken to be three times the sum of the oxide and drain junction capacitances. This delay should correspond rather directly to the delay per stage measured for ring oscillators in any given process, and is plotted along with several experimental points in Figure 9.11. It is remarkable that, despite the reduction in supply voltage at small feature sizes, logic performance continues to improve. Several authors have emphasized the improvement in speed that we can make available by reducing threshold and power-supply voltages [12, 31–33].

The primary effect behind this somewhat counterintuitive trend is velocity saturation, an excellent recent account of which can be found in Noor Mohammad [29]. We gave an early treatment of the effect of velocity saturation on device characteristics [34]; an extended analysis appears in Appendix B of a previous work [2].

The supply voltage V affects the performance of standard CMOS digital logic in three ways:

1. The channel charge is proportional to $V - V_t$.
2. The electric field in the channel is proportional to V.
3. The logic swing is proportional to V.

For long-channel devices, the carrier velocity is proportional to the electric field in the channel. The channel current is the product of the channel charge and the carrier velocity. Therefore, the device current has a quadratic dependence on the

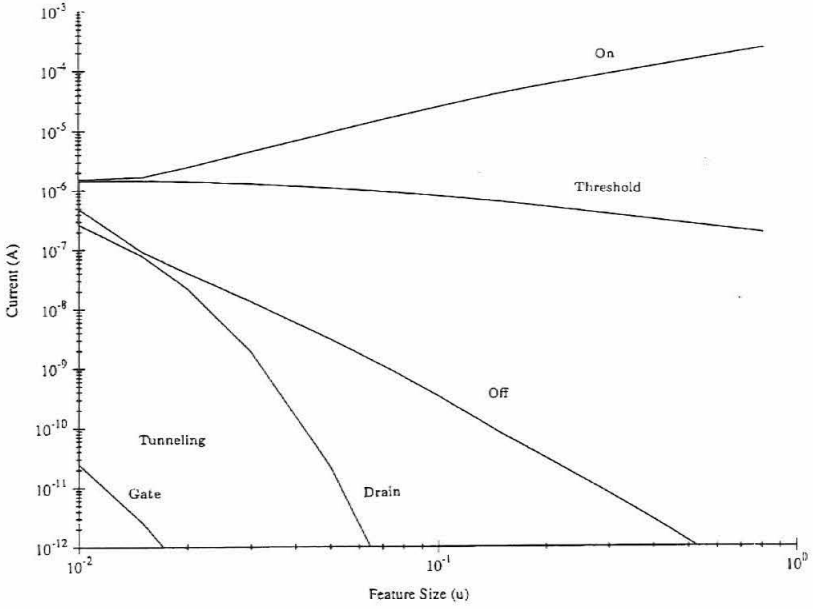

Fig. 9.9. Currents characteristic of minimum-sized devices as a function of feature size. We obtain the threshold current I_t by substituting $Q_s = Q_t$ from Equation 9.13 into Equation 9.16. We obtain the on current I_{on} by substituting Q_s from Equation 9.14 into Equation 9.16, using the threshold voltage lowered only by the built-in junction voltage, rather than by the total junction voltage. We obtain the off current Ioff by substituting Q_s from Equation 9.15 into Equation 9.16, using the threshold voltage as lowered by the full supply voltage. The junction tunneling current was computed from Equation 9.10, assuming the drain area is the square of the feature size. The gate-oxide tunneling current was computed from Equation 9.7, assuming that the full supply voltage is present across the full gate area (the square of the feature size).

supply voltage. This current must charge the load capacitance to approximately one-half of the supply voltage to achieve a logic transition. This factor cancels one of the V terms in the current, leaving the circuit speed linear in the supply voltage.

Once the carrier velocity is saturated, however, increasing the electric field in the channel no longer increases the channel current. Both the charge in transit and the voltage to be traversed by the output are increased by the same factor. In this regime, the only effect of increased supply voltage is an increase in the switching energy, with virtually no increase in performance. Just how close devices of the present day come to this limit can be seen in the delay-versus-voltage plots in the recent literature; see, for example, [6, 10, 14].

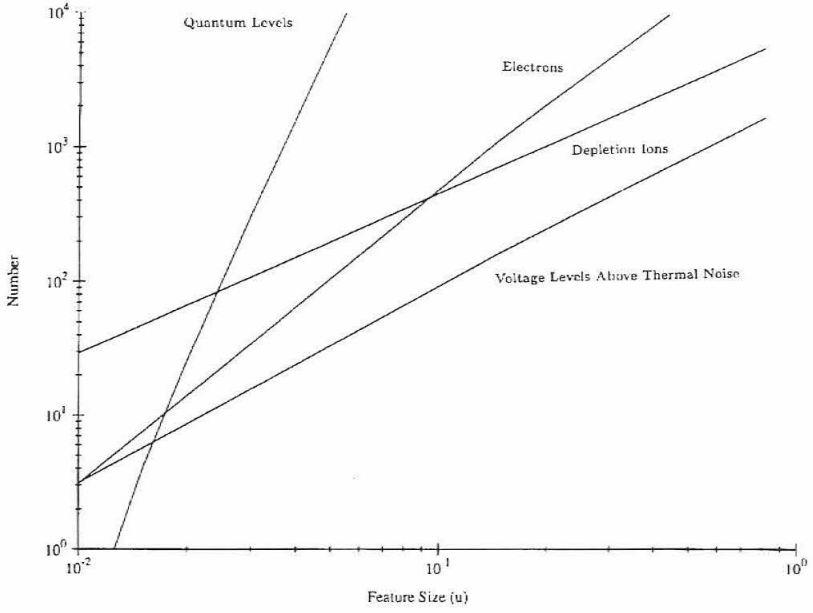

Fig. 9.10. Number of signal levels resolvable by a minimum-sized device according to the scaling laws used in this study. Thermal noise limits the analog depth representable by a single voltage. The number of voltage levels above thermal noise was taken to be the square root of the minimum stored energy shown in figure 3, expressed in units of kT. The quantum-level separation was taken to be the energy spacing of states in a one-dimensional box of length $l - x_s - x_d$. The number of electrons under the gate was taken to be the on-value of Qs multiplied by the gate area (a slight overestimate). The number of depletion ions was taken to be the doping density n given by Equation 9.8, multiplied by the gate area and the depletion depth x from Equation 9.9, using 1 V for ψ. As the number of depletion ions becomes smaller, the range of threshold voltages encountered across a single chip increases. In analog systems, adaptation techniques can mitigate or eliminate the variation among transistors.

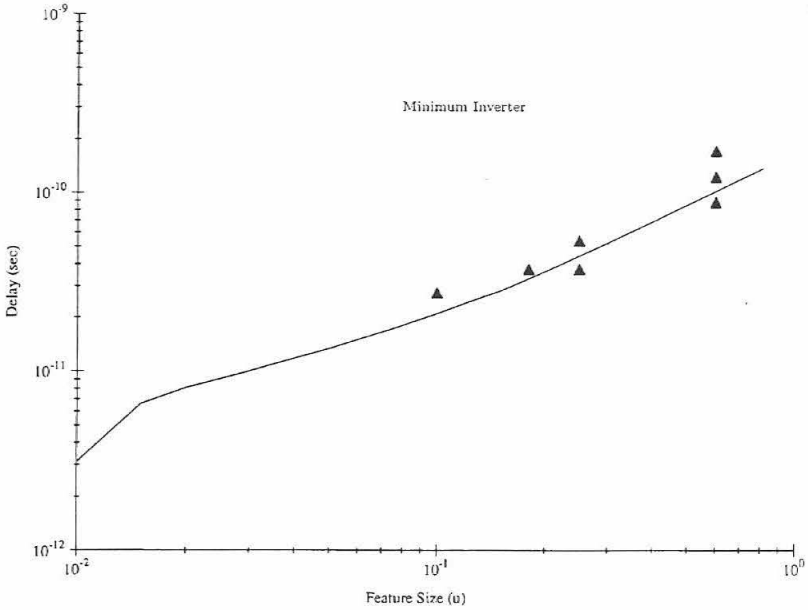

Fig. 9.11. Delay of minimally loaded inverter as a function of feature size. Filled triangles are experimental results from ring oscillators reported in the literature. Solid line is the expression given in Equation 9.17.

Because we have at our disposal the currents associated with all terminals of the transistor, we can evaluate the conductances associated with these currents. For logic devices to function properly, it is necessary that an elementary logic circuit have a gain greater than unity, which in turn requires that the transconductance G_m of the transistor be larger than the sum of all contributions to the drain conductance. As feature size decreases below 0.1 μ, both DIBL and drain-junction tunneling make rapidly increasing contributions to the drain conductance, as can be seen in Figure 9.12. In spite of these parasitic effects, the device is still capable of providing greater than unity gain down to the smallest feature sizes investigated.

9.5 System Properties

The enormous effect of device scaling on computational capability becomes apparent only when viewed from the system level. We can estimate the system-level capabilities of digital chips fabricated with advanced processes by extrapolation from present-day systems. The first such extrapolation is the number of devices per unit area. If every transistor in a modern digital chip were to be shrunk to minimum size, the entire active area would cover approximately 2% of the chip area. If we

assume that this coverage factor can be maintained in future designs, the density of active elements scales with feature size, as shown in Figure 9.13. The system clock period in today's processors is approximately 100τ. Even today, it is becoming more economical to break each chip into several processors that can operate in parallel, than it is to merely build larger "dinosaur" processors. For purposes of extrapolation, we can assume that each processor contains 106 transistors. The computation available under these clearly oversimplified assumptions is plotted versus feature size in Figure 9.14. If we further assume that all devices are in fact of minimum size, and that they are clocked at the system-clock frequency, we can estimate the power that will be dissipated by chips built in these advanced technologies. The power attributable to useful switching, and the dissipations of various parasitic currents that do not depend on clock speed, are shown in Figure 9.15. Down to about 0.03 μ feature size, most of the energy supplied to the chip is dissipated in real, useful computation. Only below this scale do the parasitic currents overwhelm the energy consumed in performing real computation.

9.6 Conclusions

The MOS transistor has become the workhorse of modern microelectronics; it has survived many generations of process scaling to finer feature sizes. In this study, I have explored the extent to which the MOS device, as we know it today, can be scaled to still smaller dimensions. We have data available to provide experimental support for the tunneling currents that will be encountered in the heavily doped source and drain junctions of devices down to 0.03 μ. Neither do we have comparable data to support the theory for oxides in the 10 $\mathring{A}$ range, nor do we have direct experimental verification of the effect of statistical fluctuations on very small structures built in heavily doped material. As such data become available, we will be better able to chart the course of future minification, of which the present study is only an outline. It is already clear that MOS circuits will be integrated to upward of 10^9 devices per square centimeter merely by scaling, without any major change in the conceptual framework that we use today. There are many challenges involved in this technology evolution [4], but I do not expect any show-stoppers. The prospect of very high levels of integration was daunting in 1971 when our earlier study was written, and is far more daunting today. Whereas massive parallelism is possible in present-day technology, it will clearly become mandatory if we are to realize even a fraction of the potential of more highly evolved technology. Even as this study is written, there is far more potential in a square centimeter of silicon than we have developed the paradigms to use, as has often been the case in periods of rapid technological evolution.

I should clarify the "limits" considered in this study. It is clear that devices much smaller than those treated here can be made to show useful characteristics. Conventional MOS devices can be fabricated on insulating substrates (SOI-SOS),

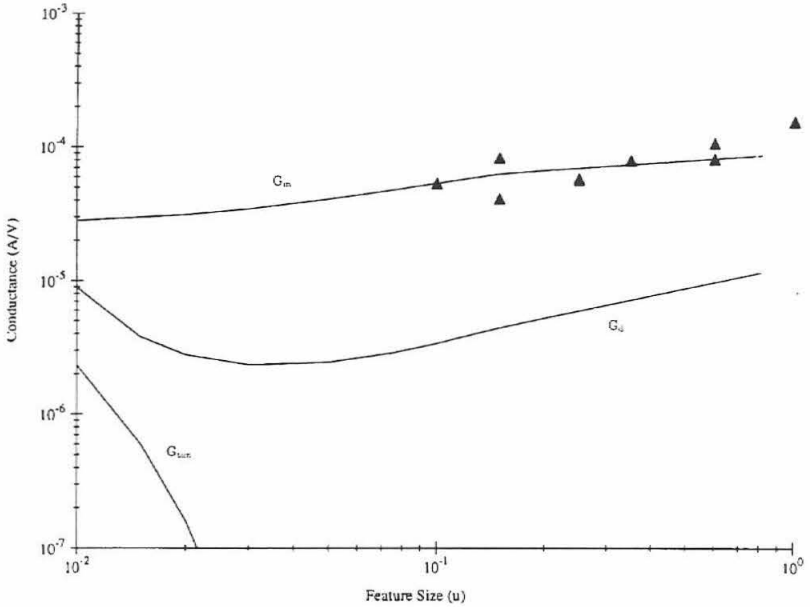

Fig. 9.12. Several conductances associated with minimum-sized transistors, as a function of feature size. The top curve is the transconductance. The filled triangles are experimental values given in the literature, normalized to a minimum-sized device at the reported dimension. The second curve is the drain conductance due to DIBL, computed by evaluating Equation 9.16 at a drain voltage equal to V and at 0.9 V, and dividing the difference by 0.1 V. The current through this conductance flows from drain to source. The bottom curve is the drain conductance due to drain junction tunneling. Current through this conductance flows from drain to substrate.

thereby removing the constraint imposed by substrate tunneling. Much smaller devices are possible at molecular scale. The most obvious example of an extremely small device is an electron-transfer reaction occurring along an amino acid path, the potential of which is determined by the charge on a nearby atomic site. Such arrangements are thought to occur in many biological systems. The physics of such a transfer corresponds directly to that of an MOS transistor operating in weak inversion (below threshold). Imagining a device that functions is easy; building a device that works is much harder; and having a process by which billions of devices can be constructed in a single physical structure is many orders of magnitude harder still. I have limited this study to the consideration of direct extensions to existing technology.

Finally, I emphasize that I have considered only the properties of transistors themselves, and have not even touched many other important aspects of the tech-

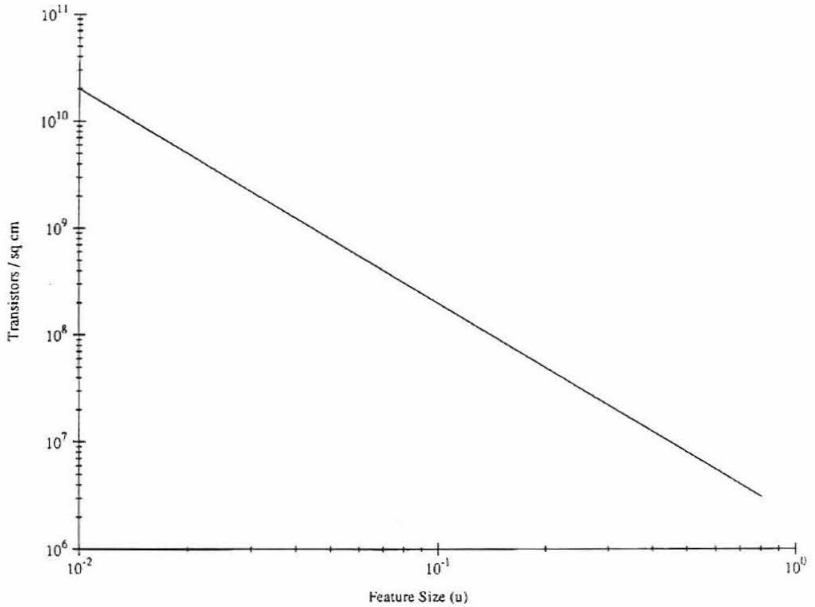

Fig. 9.13. Assumed number of active devices per square centimeter of chip area. If all devices are of minimum size, active (transistor channel) area is 2% of total area.

nology. Of the latter, interconnect — both within a single chip and across chip boundaries — is obviously a key concern. We have given elsewhere preliminary discussion of the global scaling properties of a single-chip interconnect network for ultradense technology [1]. The topic of interconnect, along with many other issues, such as the fabrication technology itself, deserve a great deal of consideration as the technology evolves. Whatever complications arise, however, it is clear that the technology will evolve. It will evolve because that evolution is possible, because there is so much to be gained at the system level by that evolution, and because the same energy and will on the part of bright, energetic, devoted people that has overcome enormous obstacles in the past will overcome those that lie ahead.

References

[1] C. Mead and L. Conway, *Introduction to VLSI Systems* (Addison-Wesley: Reading, MA, 1980).

[2] C. Mead, *Analog VLSI and Neural Systems* (Addison-Wesley: Reading, MA, 1989).

[3] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics. I: MOS technology," *Solid-State Electron.*, Vol. **15**, pp.819-829 (1972).

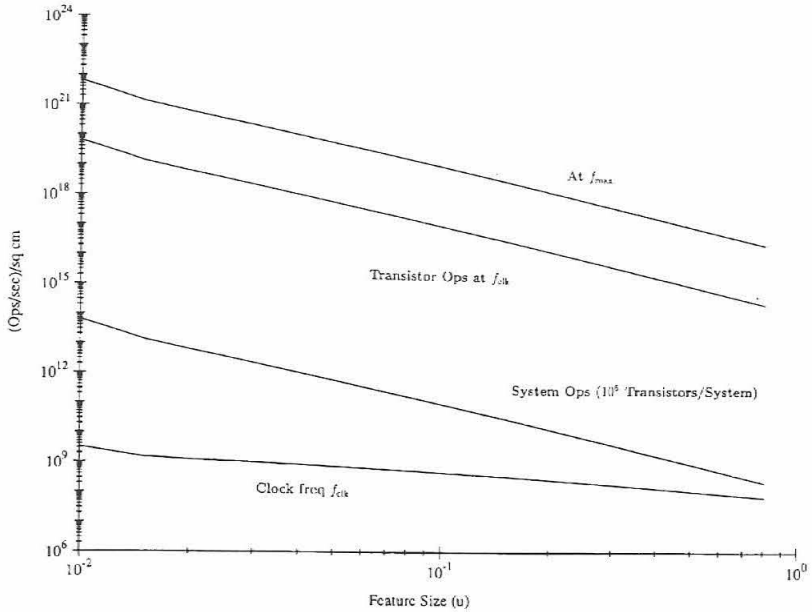

Fig. 9.14. Several measures of computation capability per unit area as a function of feature size. The bottom curve is a typical processor clock frequency, the clock period assumed to be 100 times the inverter delay shown in Figure 9.11. The second curve is the number of systems (of 10^6 transistors each) per square centimeter multiplied by the clock frequency. The third curve is the number of transistors per square centimeter shown in Figure 9.13 multiplied by the clock frequency. The top curve is the number of transistors per square centimeter multiplied by the reciprocal of the inverter delay shown in Figure 9.11.

[4] M. Nagata, "Limitations, innovations, and challenges of circuits and devices into a half micrometer and beyond," *IEEE J Solid-State Circuits*, Vol.**27**, pp.465-472 (1992).

[5] B. Davari, W.-H. Chang, K. F. Petrillo, CY. Wong, D. May, Y. Taur, M. R. Wordeman, J. Y-C. Sun, C. C-H. Hsu, and M. R. Polcari, "A high-performance 0.25-μm CMOS technology. II: Technology," *IEEE Thans. Electron Dev.*, Vol.**39**, pp.967-975 (1992).

[6] W.-H. Chang, B. Davari, M. Wordeman, Y. Taur, CC-H. Hsu, and M. D. Rodriguez, "A high-performance 0.25-μm CMOS technology. I: Design and characterization," *IEEE Trans. Electron Dev.*, Vol.**39**, pp.959-966 (1992).

[7] A. Bryant, B. EI-Kareh, T. Furukawa, W. P Noble, E. J. Nowak, W. Schwittek, and W. Tonti, "A fundamental performance limit of optimized 3.3-V sub-quarter-micrometer fully overlapped LDD MOSFETs," *IEEE Trans. Electron Dev.*, Vol.**39**, pp.1208-1215 (1992).

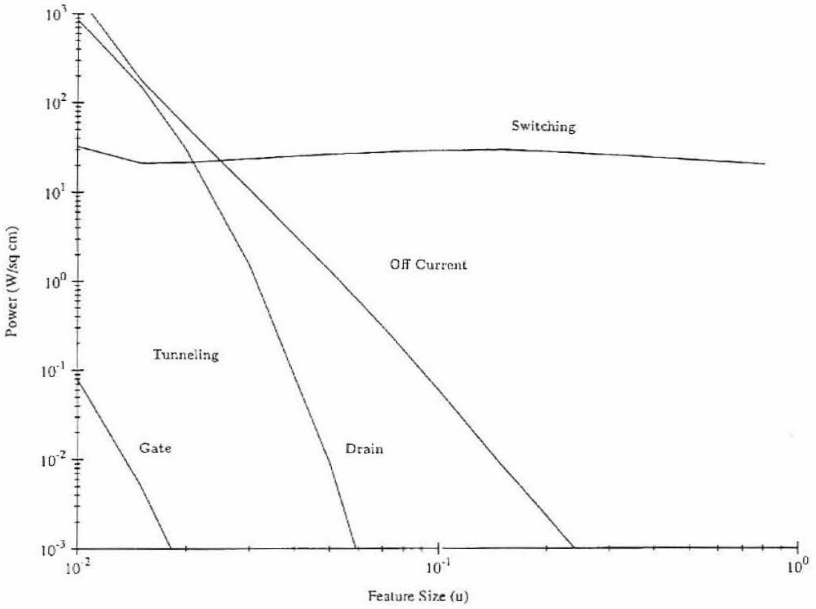

Fig. 9.15. Several contributions to the power dissipated by typical digital systems as a function of feature size. I obtained the curve labeled **Switching** by multiplying the number of transistors per unit area shown in Figure 9.13 by the switching energy shown in Figure 9.3 and by the clock frequency shown in Figure 9.14. This power contributes to the performance of computation: It scales directly with clock frequency. In addition to the switching power, there are several parasitic mechanisms by which power is wasted, each being the result of one of the parasitic currents shown in Figure 9.9. These parasitic mechanisms are present even at zero clock frequency, and perform no useful work. The values shown assume that all devices are of minimum size, and have the full voltage V on their drains. All values depend critically on the assumptions embodied in the scaling laws of Equations 9.1, 9.2, 9.8, and 9.11. Even slightly different scaling can lead to substantially different results for the smallest feature sizes. The particular laws that I have put forth in this study were fine tuned to produce reasonable results down to 0.02 μ. For example, a slight increase in doping density markedly decreases the off current by reducing DIBL, while dramatically increasing the drain-junction tunneling current. Similar trade-offs can be made with other parameters.

[8] R. H. Yan, K. E Lee, DY Jeon, YO. Kim, D. M. Tennant, E. H. Westerwick, G. M. Chin, M. D. Morris, K. Early, and P Mulgren, "High-performance deep-submicrometer Si MOSFETs using vertical doping engineering," *IEEE Trans. Electron Dev.*, Vol.**39**, p.2639 (1992).

[9] Y. Yamaguchi, A. Ishibashi, M. Shimizu, T. Nishimura, K. Tsukamoto, K. Horie, and Y. Akasaka, "A high-speed 0.6-μm 16K CMOS gate array on a thin SIMOX film," *IEEE Trans. Electron Dev.*, Vol.**1**, pp.179-185 (1993).

[10] M. Iwase, T. Mizuno, M. Takahashi, H. Niiyama, M. Fukumoto, K Ishida, S. Inaba, Y. Takigami, A. Sanda, A. Toriumi, and M. Yashimi, "High-performance 0.10-μm CMOS devices operating at room temperature," *IEEE Electron Dev. Lett.*, Vol.**14**, pp.51-53 (1993).

[11] C. Hu, "Future CMOS scaling and reliability," *Proc. IEEE*, Vol. **81**, pp.682-689 (1993).

[12] R. F Lyon, "Cost, power, and parallelism in speech signal processing," in *Proc. IEEE 1993 Custom Integrated Circuits Conf*, San Diego, CA, pp. 15.1.1-15.1.9, 1993.

[13] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO2," *J. Appl. Phys.*, Vol.**40**, pp.278-283 (1969).

[14] T. Hori S. Akamatsu, and Y. Odake, "Deep-submicrometer CMOS technology with reoxidized or annealed nitrided-oxide gate dielectrics prepared by rapid thermal processing," *IEEE Trans. Electron Dev*, Vol.**39**, pp. 118-126 (1992).

[15] J. Suñé, P. Olivo, and B. Riccò, "Quantum-mechanical modeling of accumulation layers in MOS structure," *IEEE Trans. Electron Dev.*, Vol.**39**, pp. 1732-1739 (1992).

[16] A. G. Chynoweth, W. L. Feldmann, C. A. Lee, R. A. Logan, and G. L. Pearson, "Internal field emission at narrow silicon and germanium $p - n$ junctions," *Phys. Rev*, Vol.**118**, pp.425-434 (1960).

[17] R. A. Logan and A. G. Chynoweth, "Effect of degenerate semiconductor band structure on current-voltage characteristics of silicon tunnel diodes," *Phys. Rev.*, Vol. **131**, pp.89-95 (1963).

[18] J. B. Krieger, "Theory of electron tunneling in semiconductors with degenerate band structure," *Ann. Phys.*, Vol.**36**, pp.1-60 (1966).

[19] R. B. Fair and R. W. Wivell, "Zener and avalanche breakdown in as-implanted low-voltage Si n-p junctions," *IEEE Trans. Electron Dev*, Vol. **ED-23**, pp. 512-518 (1976).

[20] J. M. C. Stork and R. D. Isaac, "Tunneling in base-emitter junctions," *IEEE Trans. Electron Devi.*, Vol. **ED-30**, pp. 1527-1534 (1983).

[21] F. Hackbarth and D.-L. Tang, "Inherent and stress-induced leakage in heavily doped silicon junctions," *IEEE Trans. Electron Dev.*, Vol. **35**, pp. 2108-2118 (1988).

[22] M. Reisch, "Tunneling-induced leakage currents in *pn* junctions," AEÜ, Band **44**, pp.368-376 (1990).

[23] G. P Li, F. Hackbarth, and T.-C. Chen, "Identification and implication of a perimeter tunneling current component in advanced self-aligned bipolar transistors," *IEEE Trans. Electron Dev*, Vol. **ED-35**, pp.89-95 (1988).

[24] R. R. Troutman, "VLSI limitations from drain-induced barrier lowering," *IEEE Trans. Electron Dev.*, Vol. **ED-26**, pp.461-469 (1979).

[25] M. J. Deen and Z. X. Yan, "DIBL in short-channel NMOS devices at 77 K," *IEEE Trans. Electron Dev*, Vol.**39**, pp.908-915 (1992).

[26] M. J. Van der Tol and S. G. Chamberlain, "Drain-induced barrier lowering in buried-channel MOSFETs," *IEEE Trans. Electron Dev*, Vol. **40**, pp.741-749 (1993).

[27] J. G. C. Bakker, "Simple analytical expressions for the fringing field and fringing-field-induced transfer time in charge-coupled devices," *IEEE Trans. Electron Dev.*, Vol. **38**, pp.1152-1161 (1991).

[28] T. A. Fjeldly and M. Shur, "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs," *IEEE Trans. Electron Dev.*, Vol.**40**, pp.137–145 (1993).

[29] S. Noor Mohammad, "Unified model for drift velocities of electrons and holes in semiconductors as a function of temperature and electric field," *Solid-State Electron.*, Vol.**35**, pp. 1391-1396 (1992).

[30] K. Nishinohara, N. Shigyo, and T. Wada, "Effects of microscopic fluctuations in dopant distributions on MOSFET threshold voltage," *IEEE Trans. Electron Dev.*, Vol.**39**, pp.634-639 (1992).

[31] J. B. Burr and A. M. Peterson, "Energy considerations in multichip-module based multiprocessors," in *IEEE Int. Conf, Computer Design*, pp.593-600 (1991).

[32] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages," *IEEE J. Solid-State Circuits*, Vol.**28**, pp.10-17 (1993).

[33] B. T. Murphy, "Minimization of transistor delay at a given power density," *IEEE Trans. Electron Dev.*, Vol.**40**, 414-420 (1993).

[34] B. Hoeneisen and C. A. Mead, "Current-voltage characteristics of small size MOS transistors," *IEEE Trans. Electron Dev.*, Vol. **19**, pp.382-383 (1972).